14

Language Documentation & Conservation Special Publication No. 15
Reflections on Language Documentation 20 Years after Himmelmann 1998
ed. by Bradley McDonnell, Andrea L. Berez-Kroeker & Gary Holton, pp. 140–149
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/24821

Reflections on software and technology for language documentation

Alexandre Arkhipov
University of Hamburg
Lomonosov Moscow State University

Nick Thieberger
University of Melbourne

Technological developments in the last decades enabled an unprecedented growth in volumes and quality of collected language data. Emerging challenges include ensuring the longevity of the records, making them accessible and reusable for fellow researchers as well as for the speech communities. These records are robust research data on which verifiable claims can be based and on which future research can be built, and are the basis for revitalization of cultural practices, including language and music performance. Recording, storage and analysis technologies become more lightweight and portable, allowing language speakers to actively participate in documentation activities. This also results in growing needs for training and support, and thus more interaction and collaboration between linguists, developers and speakers. Both cutting-edge speech technologies and crowdsourcing methods can be effectively used to overcome bottlenecks between different stages of analysis. While the endeavour to develop a single all-purpose integrated workbench for documentary linguists may not be achievable, investing in robust open interchange formats that can be accessed and enriched by independent pieces of software seems more promising for the near future.

1. Introduction[1] A major factor in the rise of language documentation (LD) since Himmelmann (1998) has been the move to digital methods, allowing an increase in

[1]We thank the editors of this volume for their encouragement, and two anonymous reviewers for their valuable comments. All errors and possible shortcomings are our own. The contribution by A. Arkhipov has been made in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies

ISBN: 978-0-9973295-3-7

recordings and in the amount transcribed, and providing for analyses more firmly based in citable data than was previously the case.

The landscape for computer-assisted linguistic processing has changed considerably over the past 20 years. Fieldwork techniques have dramatically improved, with lightweight video and audio recording equipment, and new tools for annotating, analyzing and archiving language data. Further, with the emphasis on archiving has come better data management and methods for delivery of language records back to the source communities, not something that was easily or commonly done with analog recordings. Making these recordings available exemplifies the responsibility of academic researchers to work collaboratively and to ensure reusability of their data. As materials go back to the source community so also are new materials being created by members of that community.

Technology allows distant collaborations, and also allows for research to have multiple outputs from well-structured primary and secondary data and annotations. Differential access to language records between researchers and language speakers has dramatically reduced in many places and is rapidly shrinking in the rest. Mobile devices are making it easier to create and disseminate records of performance in small languages and social media promotes interactions in local languages. Digital media can be made available in formats derived from higher resolution archival formats, with descriptions that provide contextual information describing what is in the media. Ideally, transcripts of the contents of the media are also available and allow users to locate targeted points within cultural records.

However, the use of technology in recording language performance is not, in itself, sufficient to ensure the quality of the recording, nor to ensure its longevity. Thus, for instance, the position of a microphone is crucial to ensuring a good-quality recording, as is the use of a windscreen in outdoor settings. Analog recordings made on magnetic tape are now nearing the end of their playable life so digitization of the existing legacy of language recordings is one of today's urgent priorities. The standards specified by the international community of sound archives[2] need to be understood by linguistic researchers and applied routinely. As for the written domain, there was initial delight in being able to create complex documents like grammars and dictionaries using word-processors, but it soon became apparent that the proprietary formats they used could lock data away unless converted to another format.

There needs to be regular training in the use of technology so that the basic principles and standards are known and can be followed even as particular tools become obsolete. An example of a widely adopted standard is the Leipzig Glossing Rules[3] which have improved cross-linguistic interpretation of glossed texts. It should be stressed that adhering to such standards is not an onerous condition on research and can take as little as reading a brief document or attending a training workshop. On the other hand, it must be acknowledged that although basic principles are quite straightforward to master, the details of use of particular tools and interaction between tools in different setups are highly specific and can often be a source of frustration. Thus not only an effort is required from the LD practitioners to invest in learning, but considerable effort is also required from the developers to invest in harmonization of tools and making workflows more straightforward and robust.

---

of Sciences and Humanities. Thieberger is an ARC Future Fellow and a CI in the ARC Centre of Excellence for the Dynamics of Language.

[2] `https://www.iasa-web.org/tc04/audio-preservation`

[3] `https://www.eva.mpg.de/lingua/resources/glossing-rules.php`

In what follows, we start with a 20 years' flashback (§2), then proceed to review several important developments in technologies and workflows since that time (§3), and conclude with some speculative remarks on the future of technologies in LD (§4).

2. Back in 1998 Looking back to 1998, perhaps the biggest changes are in the capture, use, and analysis of dynamic media. Storage was expensive and data transfer took a long time. The first USB drives appeared in 2000 and the first terabyte disk in 2007.[4] For audio recording equipment there was a choice between cassette recorders, minidisc (1992–2013), or DAT (1987–2005). Digital video had been available since 1986, using digital cassette tapes.

There were no simple transcription tools: ELAN[5] was not yet developed, Transcriber[6] was first released in 1998. In the same year, a hardware transcriber—a cassette player equipped with a pedal to repeatedly playback a portion of the tape—was the most sophisticated transcription device that Arkhipov saw used in the team fieldtrip of the Moscow State University. SoundIndex was an early transcription tool produced by Michel Jacobson at LACITO (Michailovsky et al. 2014) and was used in the first online presentation of text and media (now PANGLOSS). SoundIndex was used by Thieberger in his analysis of Nafsan (South Efate) and allowed his grammar to be the first to cite examples back to an archival media corpus (Thieberger 2009).

In 1998 the only digital indigenous language archive was text-based (the Aboriginal Studies Electronic Data Archive – ASEDA; see Thieberger 1995). The Archive of the Indigenous Languages of Latin America (AILLA)[7] began in 2000, the DoBeS programme in 2000 started the MPI Archive in Nijmegen (now the TLA).[8] The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)[9] began in 2003, and the Endangered Languages Archive (ELAR)[10] in 2004. There are now a number of other such archives (see the Open Language Archives Community (OLAC)[11] page or DELAMAN[12] for lists of affiliated archives). Associated with digital archiving is the widespread adoption of the metadata standards established by OLAC and the TLA. These have allowed significant information sharing about the content of archives and increased access to the records they hold.

3. Advances and emerging challenges over the past 20 years In this section, we briefly review several deliberately selected technology-related issues, necessarily leaving aside many others, by no means less important. We will touch upon ensuring longevity of LD data (§3.1), as well as upon transitioning between various analysis stages (§3.2) and tools (§3.4), going portable (§3.3), publishing LD data (§3.5) and, finally, training in technologies (§3.6).

3.1 Data preservation and management While there are many benefits of digital technology, problems have become more apparent with experience. The fragility of

[4] http://www.computerhistory.org/timeline/memory-storage
[5] https://tla.mpi.nl/tools/tla-tools/elan/
[6] http://trans.sourceforge.net/en/presentation.php
[7] https://ailla.utexas.org/
[8] https://tla.mpi.nl/
[9] http://paradisec.org.au
[10] https://www.soas.ac.uk/elar/
[11] http://www.language-archives.org/
[12] http://delaman.org

digital data means paying attention to backing up, especially in climates where equipment life is compromised by moisture. Being able to record more easily has the corollary of creating many more files to manage. Data (and metadata) management is now a serious consideration (in all research, not just in linguistics, see Corti et al. 2014) and is being taught as part of field methods courses.

Aside from the preservation and management of digital files themselves, the evolution of hardware and software requires continuous migration of data to newer formats and storage media. As software has come and gone over the past two decades it has prompted thinking about how our data can survive the tools we use. Many of us have had the experience of finding files created in the past that are no longer accessible. It is also enticing to produce a multimedia app or website that has rich content, but, again, the primary data used in these has to be kept in an open format, as multimedia products have a very short lifespan. Sustaining data means ensuring it can continue to be read and that there are copies of it in a number of locations, ideally also in an archive. To do this, it is best to keep an open format or text copy of files rather than have them stored inside a proprietary format (like Microsoft formats xls or doc).

A concomitant problem, very acute before the Unicode standard came into wide use, was competing and idiosyncratic character encodings and fonts (see also Kalish 2007 on this issue). Combining two or more character sets in a single document presented a particular challenge, such as writing a descriptive grammar in Russian with English glosses and custom transcriptions adorned with various diacritics. Years later, even if the file format is still readable by modern software, half of the characters are hard to reconstruct. Transcoding solutions such as SILConverters[13] proved to be particularly helpful in such cases.

3.2 Transcription bottleneck With this increased volume of recordings, transcription has become a major bottleneck (see Himmelmann this volume). In fact, only a fraction of the collected data ever gets transcribed, which means that even smaller data volumes end up as fully analyzed corpora.

However, LD will be able to increasingly benefit from the technologies developed for major languages, both in spoken and written form. ASR (automatic speech recognition) and related speech technologies such as forced alignment have been shown to efficiently reduce the transcription-related manual workload. Although training acoustic models used in speech recognition normally requires large volumes (sometimes hundreds of hours) of annotated data, different methods are emerging to reduce the effort of porting such systems to work effectively on smaller speech corpora (see Strunk et al., 2014; Adams et al., 2018; Johnson et al., 2018).

Another option to overcome the bottleneck is delegating the work to native speakers. The Basic Oral Language Documentation framework (BOLD; see Reiman 2010) and similar approaches suggest recording careful re-speaking of the analyzed text by (the same or another) native speaker, which can much more easily be further transcribed by linguists on their own; other kinds of oral annotations such as oral translation or comments can also be provided. SayMore is currently a tool that supports recording both oral annotations and oral translations; it is reported to be successfully used in documenting 14 languages of Nepal (Khadgi 2017).

[13] `http://scripts.sil.org/enccnvtrs`

Other initiatives support transcription crowdsourcing through an online repository where people can register to provide transcriptions for items of their choice. One such project is *Euskal Herriko Ahotsak* (Voices of the Basque Country),[14] an archive of thematic interviews with speakers of diverse varieties of Basque. Another is Phonemica,[15] a collection of stories in languages of China, where the recordings themselves are also contributed by the users and can be transcribed and translated online into Mandarin and English.

3.3 Portable solutions Another tendency of recent years, paralleling the growing performance and shrinking footprint of hardware, is the increasing demand for more lightweight and portable technological solutions. After desktop computers came laptops, by now ordinary and cherished companions of a fieldworker, then tablets and smartphones. Computers, formerly confined to pre- and post-fieldwork office use, are now indispensable throughout the field session. Portable devices and field conditions impose limitations on the software, including memory use and system performance, undesired dependencies on particular operating systems, frequent updates and connectivity. While tablets and smartphones may be unadapted to more complex analytical work, they can be a lifesaver when it comes to simpler operations which do not wait, like taking quick notes (including oral notes), collecting metadata or looking up a word in the dictionary. Higher quality devices can also be used to collect primary data, be it photos, video or audio—something which 20 years ago would require three separate and bulky analog devices.

For a long time, documenting a language was mostly seen as the linguist's domain. Nowadays, members of the speech community are not just 'contributing' to the documentation, but are frequently taking a leading role. Accordingly, the need arises to train the speakers to use linguistic equipment and software, or, more wisely, to produce tools which can easily be mastered by non-linguists. To name just a few, the dictionary collecting tool WeSay[16] and the organizer-and-transcriber SayMore[17] (both for Windows PCs), audio collecting and translating app LIG-Aikuma,[18] and Zahwa[19] app for documenting procedural knowledge like food cooking recipes (both for Android devices) have been successfully used to collect data in many remote locations across continents. Some of them are also integrated with bigger applications like FLEx[20] or ELAN, and/or offer options of preparing standardised archive submissions. Functions of oral annotations (see above) drastically lower the barrier of required user expertise.

3.4 Tool interoperability and data structures Language documentation comprises an array of diverse activities, each with its own focus and demands in data treatment. First linguistic tools that came into existence were rather specialized and each tackled a very limited portion of the workflow. For instance, the ability to transcribe directly into a digital format was a huge productivity boost by itself. However, while more and more data became produced and processed on different stages of the workflow, the transitions back and forth between transcripts, interlinear glosses, dictionaries became a new bottleneck.

[14] `https://ahotsak.eus/english/`
[15] `http://phonemica.net/`
[16] `https://software.sil.org/wesay/`
[17] `https://software.sil.org/saymore/`
[18] `https://lig-aikuma.imag.fr/` see also `http://www.aikuma.org/`
[19] `https://zahwa.aikuma.org/`
[20] `https://software.sil.org/fieldworks/`

Early tools like Toolbox[21] have determined the data structures used by linguists so that e.g. the common format for many bilingual dictionaries of small languages is the 'backslash' file (the SIL 'Standard Format').[22] These files are plain text and can be converted to new formats and be archived, although there is considerable variation in user-specific field codes and structure. The same format was used for storing interlinear texts, with the additional problem that it relied on counting whitespace characters for word-by-word alignment.

An important step towards better interoperability and sustainability was the adoption of XML (introduced in 1998)[23] either as native format or at least as an export/import format by most tools. While data structures embodied in XML documents vary greatly between different software, the availability of common standard-compliant processing tools makes it technically possible and relatively easy to convert between them. Yet the multitude of tools as well as their quick evolution is a challenge. In an LD project that Arkhipov took part in from 2006, all software elements and data formats used at different steps for transcribing, glossing, archiving and presentation changed within 3-5 years, which demanded considerable effort to maintain.

In the currently running long-term INEL project,[24] transcriptions coming from four different sources are imported into FLEx for glossing: plain text typed in from archival manuscripts, transcripts by native speakers in common office format, transcripts done in SayMore by more computer-proficient native speakers and those made in ELAN by linguists. Once glossed, the texts are exported into EXMARaLDA[25] format for further annotation and presentation. This is all possible thanks to the interaction between ELAN and FLEx which improved greatly since 2008, now preserving speaker, time-alignment and mediafile attributes crucial for time-aligned glossed text corpora. However, the inability of FLEx to import existing morpheme glosses remains a major blocker. It not only makes it impossible to incorporate external changes to any aspect of at least partly glossed text, but also prevents many from using FLEx altogether, especially those having a substantial corpus analyzed elsewhere (e.g. in Toolbox or Word). Another one is lack of support for custom annotation tiers in FLEx. These two problems are however not inherent to the FLEx interlinear XML format, which curiously is sometimes used as a pivot interchange format without accessing the FLEx application itself.

Two alternate ways of dealing with interoperability problem can be distinguished. One is adding functionality to an existing tool, thereby reducing the need to interact with other tools. Thus ELAN as primarily a multimedia transcription/annotation tool now includes a glossing module, FLEx integrates lexicon and text analysis with dictionary-publishing solutions, and SayMore combines metadata curation and file management with transcription. However, it seems that the complexity of the most sophisticated tools is close to reaching a certain limit beyond which the required development and maintenance efforts surpass the resources of the linguistic community. A hypothetical all-purpose workbench for documentary linguists would need to support a wide range of primary data including various media types, various content types (texts, words, paradigms, questionnaires, as well as metadata), flexible annotations, interlinking between text and media, complex analytical tools, rich visualization and publishing options. Not only does

[21] `https://software.sil.org/toolbox/`
[22] `http://downloads.sil.org/legacy/shoebox/MDF_2000.pdf`
[23] `https://www.w3.org/TR/1998/REC-xml-19980210`
[24] `https://inel.corpora.uni-hamburg.de/`
[25] `http://exmaralda.org/en/`

such omni-functionality require deep insight into possible use cases to be adequately designed and an extremely diverse developer expertise, but the application becomes increasingly heavy and slow which ultimately restricts possible use cases (cf. §3.3). On the other hand, a universal tool must also be cross-platform: meanwhile, developer experience tells us that in this case a lion's share of worktime is taken by making each and every feature function identically in different computational environments. These are perhaps the most evident reasons why a single all-purpose LD tool is unlikely to appear in the foreseeable future.

An alternative is, then, to ensure lossless and efficient omnidirectional data transfer between independently developed tools. In this view, open, well-documented (and possibly human-readable) interchange formats are a priority. A recent release of CLDF, a specification for Cross-Linguistic Data Formats,[26] is a promising step along this path. It proposes an interchange standard for linguistic datasets representable in tabular form. Leveraging the positive experience of a series of cross-linguistic projects such as WALS and Glottolog,[27] it explicitly aims to decouple the development of software tools from that of datasets. CLDF makes use of plain text tab-delimited files, which can be read and edited by humans and on the other hand are supported by a wide range of software. This makes the format particularly suitable for configuring custom tool chains for multi-step data processing.

3.5 Data vs. presentation It is important to present language records in a way that is accessible and attractive, but this presentation should be considered as a kind of exhibition derived from the underlying collection, not as the only product of documentation. A typical example is a lexical database in which as much information as possible is stored about each word, ideally including encyclopaedic information. Dictionaries of various kinds can be derived from this lexical database: a detailed dictionary, a learner's dictionary, or a topical dictionary. These can be presented in several ways: on paper, as a website, as an app. While a book or website can go out of date or be lost, the primary data must continue to be available for use in future. Similarly, a digital corpus can be presented as a resource for linguists, with full annotation and complex search facilities, or as a text collection with a focus on the speakers and the story, for a wider audience including the speech community.

Putting aside large archives with established infrastructure, there is currently no widely accepted and easily reusable solution for publishing LD data in a user-friendly manner. For lexical data, one should mention the no longer developed LexiquePro[28] which can generate a set of static HTML pages (for publishing on one's own website), and the current online Webonary[29] repository which allows users to register and upload their Toolbox or FLEx lexical databases for general access, providing dynamic browsing and search capabilities (although little customization). Similar attempts have been made for sharing interlinear texts, such as the Kratylos[30] project (yet rather at proof-of-concept stage). More powerful solutions, such as the recent Tsakorpus[31] corpus platform,

[26] `http://cldf.clld.org/`
[27] `http://glottolog.org/`
[28] `http://www.lexiquepro.com/`
[29] `https://www.webonary.org/`
[30] `https://www.kratylos.org/`
[31] `https://bitbucket.org/tsakorpus/`

generally require installing and configuring one's own instance of a server application which assumes an available server and substantial technical competence.

3.6 Training and support As new tools and methods appear, there is a need to train LD practitioners and to provide advice about emerging devices (cameras, recorders, microphones and so on). Such training has been provided at summer schools or training workshops, such as those run by InField/CoLang,[32] LLL,[33] ELDP,[34] or DoBeS.[35] An email list and website run by the Resource Network for Linguistic Diversity[36] has provided advice for subscribers since 2004. As software evolves even more rapidly than hardware, can be quite complex and contain poorly documented features and bugs, one-off training is usually not enough to ensure seamless work in the future. Contacting developers directly is not rarely the most efficient strategy, however not always realistic; one then has to rely on the user community and fellow researchers for continuous support and advice. Also the issue of translating the user interface and help pages should not be neglected. Here again, the LD user community is often an important actor, since developers' resources are limited.

4. Future of LD technologies Building appropriate methods into normal fieldwork practices results in records that can be archived and so made accessible to source communities. These records are robust research data on which verifiable claims can be based and on which future research can be built, and are the basis for revitalization of cultural practices, including language and music performance.

Looking ahead we can predict that recording and storage technologies will become cheaper, smaller and more intuitive, which will make it easier for more documentation by speakers, increasing the need for LD networks to train speakers and to provide ways of storing the records they create into the future. As a means of sharing these records, social networks and media platforms are likely to increasingly become centres of activity in writing, recording and distributing language performance. Crowdsourcing is starting to be used for transcribing, translating and annotating the collected data. At the same time, advanced technologies related to speech recognition, translation, text-based annotation will be increasingly applied to LD data as another remedy against the 'transcription bottleneck.'

On the other hand, harmonising, interlinking and reusing data across projects will require increased attention to develop and promote standardised software-independent formats for various data types. Such formats would allow independent tools and services to access a portion of data from a repository, process it and return enriched data which can then further be accessed by other tools and users. Provided a standard interface to access arbitrary small pieces of data, dynamic annotations similar to formulas in a spreadsheet document could become instrumental to speed up annotating large corpora and to enable updating interdependent annotations. Such architecture would also facilitate creating varied presentation formats from the same linguistic data, addressing different audiences and adapting to different devices and environments.

[32] `https://www.alaska.edu/colang2016/`
[33] `http://www.ddl.cnrs.fr/colloques/31_2012/`
[34] `http://www.eldp.net/en/our+trainings/about/`
[35] `http://dobes.mpi.nl/dobesprogramme/training_courses/`
[36] `http://rnld.org`

References

Adams, Oliver, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird & Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In Calzolari, Nicoletta (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA), 3356–3365.

Corti, Louise, Veerle van den Eynden, Libby Bishop & Matthew Woollard. 2014. *Managing and sharing research Data: A guide to good practice.* London: Sage Publications.

Johnson, Lisa M., Marianna Di Paolo & Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing Prosodylab-Aligner with Tongan data. *Language Documentation & Conservation* 12. 80–123. http://hdl.handle.net/10125/24763

Kalish, Mia. 2007. Review of Fontographer. *Language Documentation & Conservation* 1(2): 301–311. http://hdl.handle.net/10125/1723

Khadgi, Mari-Sisko. 2017. Large-scale language documentation in Nepal: A strategy based on SayMore and BOLD. Paper presented at 5th International Conference on Language Documentation and Conservation (ICLDC), Honolulu, Hawai'i, March 4, 2017. http://hdl.handle.net/10125/42029

Michailovsky, Boyd, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François, Evangelia Adamou. 2014. Documenting and researching endangered languages: The Pangloss Collection. *Language Documentation & Conservation* 8: 119–135. http://hdl.handle.net/10125/4621

Reiman, D. Will. 2010. Basic oral language documentation. *Language Documentation & Conservation* 4. 254–268. http://hdl.handle.net/10125/4479

Strunk, Jan, Florian Schiel & Frank Seifart. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation.* Reykjavik, Iceland: European Language Resources Association (ELRA), 3940–3947. http://www.lrec-conf.org/proceedings/lrec2014/summaries/1176.html

Thieberger, Nick. 1995. The Aboriginal Studies Electronic Data Archive, *International Journal on the Sociology of Language* 113. 147–150.

Thieberger, Nick. 2006, Computers in Field Linguistics. In Keith Brown (ed.), *Encyclopedia of language & linguistics*, 2nd edn., vol. 2, 780–783. Oxford: Elsevier. http://hdl.handle.net/11343/34940

Thieberger, Nick. 2009. Steps toward a grammar embedded in data. In Patience Epps & Alexandre Arkhipov (eds.), *New challenges in typology: Transcending the borders and refining the distinctions*, 389–408. Berlin: Mouton de Gruyter. `http://repository.unimelb.edu.au/10187/4864`

Alexandre Arkhipov
alexandre.arkhipov@uni-hamburg.de

Nick Thieberger
thien@unimelb.edu.au
orcid.org/0000-0001-8797-1018